

Reviewer #2

General Evaluation

This manuscript addresses a very timely, practical, and important topic for the geoscientific and geographical community. The misuse and over-interpretation of statistical significance maps (and multiple-testing corrections) is a widespread issue, and the paper clearly highlights key pitfalls that researchers encounter when applying False Discovery Rate (FDR) corrections to spatially coherent fields. Overall, the study is well-motivated and provides a valuable discussion.

However, I have several major conceptual points and specific recommendations that I believe should be addressed to strengthen the manuscript and ensure statistical accuracy before publication.

We thank the reviewer for the thorough and helpful comments. Below, we respond to each comment individually.

Major Comments

1. Inherent Nature of FDR Control vs. FWER Control

The authors should more explicitly clarify the fundamental mechanism of False Discovery Rate (FDR) control. By definition, controlling FDR means that the threshold adjusts dynamically depending on the proportion of rejections: it permits identifying more false rejections when the analysis contains many rejections (percentage-wise), and fewer false rejections when only a small fraction of tests are rejected. This property directly explains why restricting the analysis domain to a region with widespread anomalies results in a high critical threshold ($p^* = 9.8\%$), whereas analyzing the global domain (where significant tests constitute a small percentage) yields a strict threshold ($p^* = 0.99\%$).

If a researcher finds this adaptive behavior undesirable or problematic for their application, FDR control is simply not the appropriate tool. In such cases, Family-Wise Error Rate (FWER) control should be recommended, as it strictly guards against any false rejection and eliminates these domain-dependent threshold fluctuations.

We have clarified the mechanism of FDR control as suggested. We now explain that this property is intrinsic to FDR but seldom acknowledged when applying FDR corrections, and that applications in the geosciences literature often implicitly or explicitly assume that FDR makes the test more stringent.

We agree with the reviewer that FWER-controlling methods may be an alternative in these cases. However, FWER-controlling methods derived under an independence assumption tend to become overly conservative under positive spatial correlation, achieving a true error rate much smaller than the nominal local significance level, and an

under-detection of signals (Fig. R1; Livezey and Chen, 1983). This makes employing FWER-controlling methods less desirable when analyzing geospatial data where positive spatial correlation is common. While we suggest a modified expression of the Walker criterion (Walker, 1914) that takes spatial correlation into account, the primary aim of our manuscript is not to derive a new method for FDR control. Nevertheless, some type of FWER control may be a potential alternative, so we now discuss the potential use of FWER, its advantages and limitations, in Sect. 6 when presenting practical recommendations.

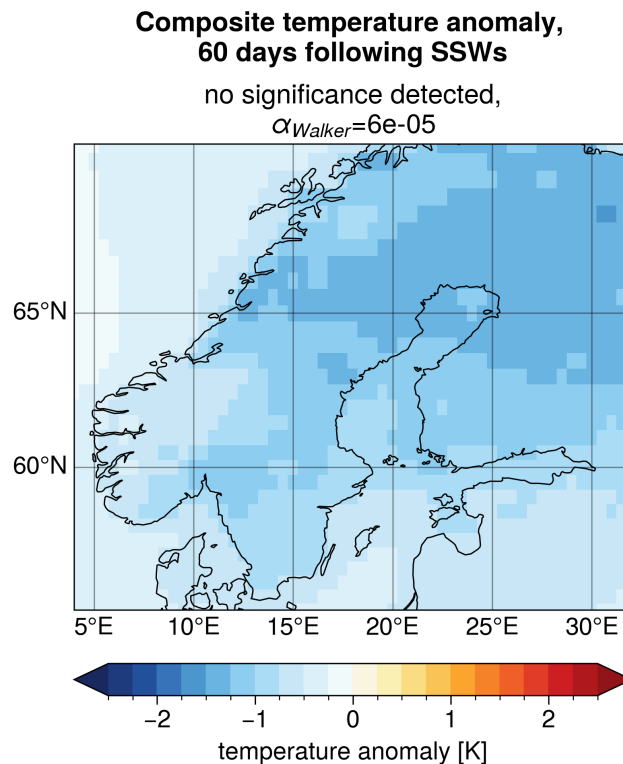


Figure R1: Composites of 2-m air temperature anomalies across northern Europe averaged over 60 days following SSWs. Statistical significance was assessed following the Walker criterion (Walker, 1914): points were considered statistically significant if the respective p-value at a grid point is below $\alpha_{walker} = 1-(1-\alpha)^{(1/N)}$ with $\alpha=0.1$ and $N=1767$, considering only grid points over northern Europe.

2. Clarification on FDR Threshold vs. Pointwise Significance (Alpha Selection)

Throughout the text, it is mentioned that FDR control can lead to a higher adjusted threshold than nominal pointwise testing (and thus reject more hypotheses). It is crucial to clarify that under identical nominal significance levels ($\alpha_{FDR} = \alpha_{pointwise}$), the FDR threshold p^* can never exceed $\alpha_{pointwise}$ (since $p_{(i)} \leq (i/N) * \alpha \leq \alpha$).

The observed phenomenon where $p^* > 0.05$ occurs solely because two different nominal levels were chosen ($\alpha_{pointwise} = 0.05$ vs. $\alpha_{FDR} = 0.1$, following Wilks,

2016). Therefore, this is not a statistical failure or anomaly of the FDR procedure itself, but rather a direct artifact of choosing a higher initial alpha for FDR. This distinction must be made explicit throughout the manuscript.

We thank the reviewer for raising this important point. We have made adjustments in line with the reviewer's specific comments below.

We nonetheless chose to retain $\alpha_{\text{FDR}}=0.1$ because this particular value for the false discovery rate has been employed in hundreds of papers in the geosciences since first suggested in the influential work by Wilks (2016). The rationale behind this choice of α_{FDR} in Wilks (2016) is that it would approach a global test level of $\alpha_{\text{global}} = 0.05$ for moderate to strong spatial correlation (Discussion of Fig. 4 in Wilks, 2016). Results may be reasonable for domains, where the ratio of locally significant points at the 5% level is relatively low, but we find that it does not directly depend on the spatial correlation, but rather the ratio of locally significant points to all grid points considered. We do not wish to imply that this is a failure of the FDR algorithm, but rather to underscore that it is a property that is seldom understood or acknowledged when the FDR correction is applied in the geosciences literature. Furthermore, α_{FDR} defines the highest allowed proportion of false positives (e.g., no more than 10% of the rejected nulls are false positives for $\alpha_{\text{FDR}} = 0.1$), and as such should not be interpreted as a local measure of statistical significance. We highlight this aspect in Sect. 2.

3. Domain Selection, Post-Hoc Analysis, and the Role of Total Test Count (N)

While the authors attribute the threshold shift across domains primarily to p-value ranking, the main driver is the reduction in total test count N relative to the proportion of small p-values. Selecting a sub-domain after visually inspecting the data represents post-hoc selection ("data fishing").

We fully agree with the reviewer. We presented this example in the paper to illustrate a common pitfall, not as an example of good scientific and statistical practice. Avoiding this post-hoc selection is already one point in our recommendations table, and something that we have tried to further emphasise in our revised manuscript following this point.

In other fields facing similar spatial/high-dimensional multiplicity problems (such as genetics), when testing a post-hoc selected region, standard practice dictates dividing by the total number of candidate tests (N) from the initial dataset, rather than the subset size n. Because the researcher selects the domain based on prior knowledge of the data, the rest of the domain is implicitly assumed non-significant. If the authors keep N equal to the global domain size when evaluating the subset, the FDR threshold for the selected domain will remain consistent with the full domain.

I strongly encourage the authors to emphasize to the geoscientific community that post-hoc spatial domain selection without retaining the full N is statistically

invalid—analogous to selecting only small values in group A and large values in group B in a two-sample test and declaring a significant difference.

We appreciate the reviewer's suggestion. However, we note that the assumption that the rest of the domain is non-significant often does not hold in geosciences applications. If there are many nominally significant gridpoints outside the selected sub-domain (e.g., at the point-wise 5%-level), as is often the case for example in the atmospheric sciences, using the full N in the FDR-correction would result in a stricter threshold than performing the test for all grid points of the global domain. In the example of t2m over northern Europe following SSWs, considering for N all gridpoints of the northern hemisphere north of 20°N but only the t2m signal of points within the northern European domain would result in no statistical significance being detected (Figure R2). We therefore believe that, for geosciences applications, it is more robust to recommend not to perform post-hoc sub-domain selection altogether.

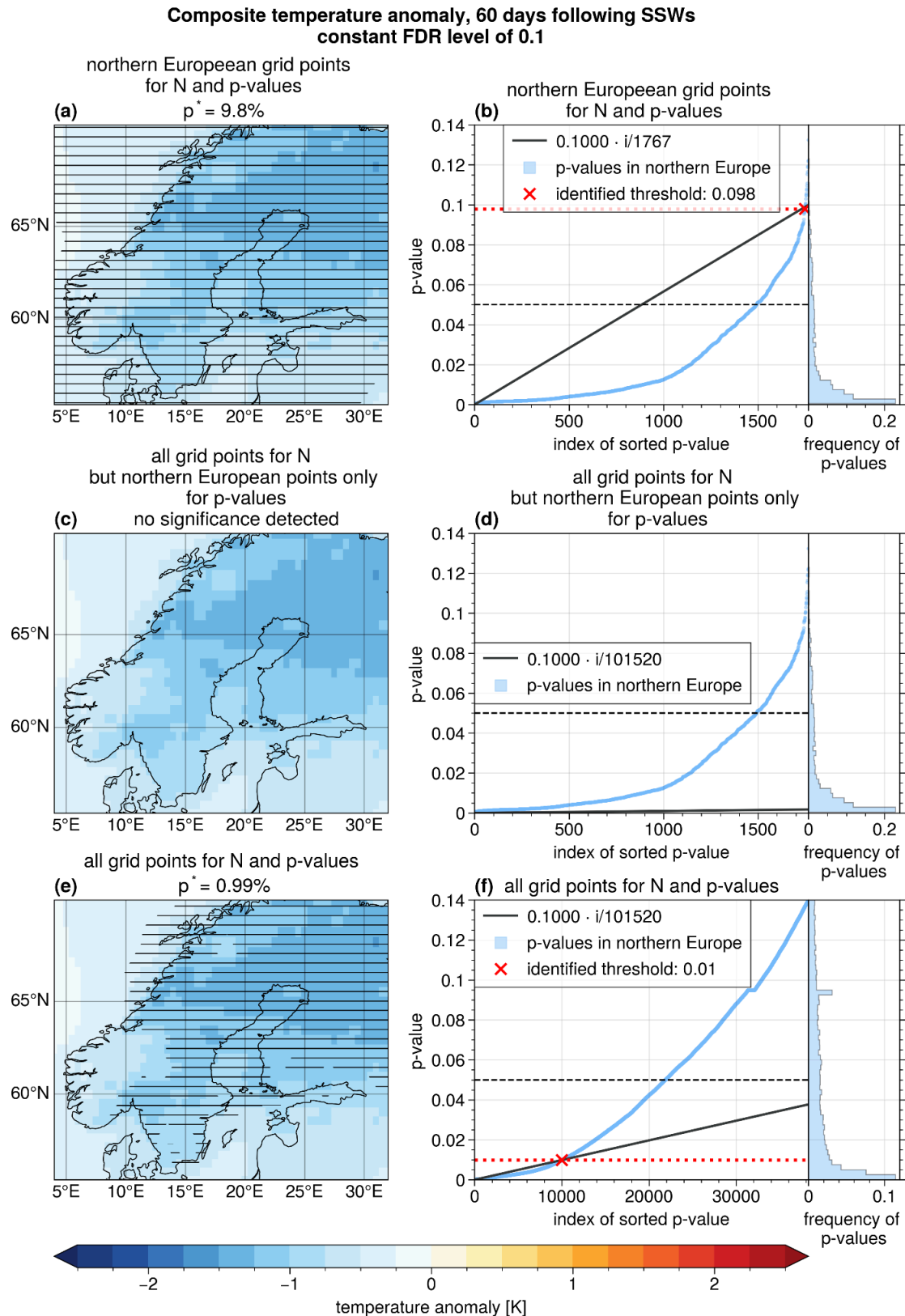


Figure R2: Left column: Composites of 2-m air temperature anomalies across northern Europe averaged over 60 days following SSWs. Right column: Ranked p-values (colored line) and the FDR threshold (black line). The first row employed significance assessment with FDR correction, corresponding to Figure 1d and 2b in the manuscript (considering

only N=1767 grid points across northern Europe), while the second row considered all gridpoints of the North Hemisphere north of 20°N in the FDR correction (N=101520) but only the p-values across northern Europe, as suggested in the reviewer's comment. The third row considered all grid points and p-values of the North Hemisphere north of 20°N, corresponding to Figure 1c and 2a in the manuscript.

4. Spatial Dependence and N_{eff} vs. Correlated Multiple Testing Procedures

Regarding the effective sample size estimation (N_{eff} in Eq. 2), substituting N_{eff} fundamentally alters the nature of the test (turning it into a Walker-type field significance test rather than FDR control). To properly handle spatial autocorrelation within the FDR framework, I suggest referring to multiple testing corrections specifically designed to account for test dependence (e.g., Benjamini & Yekutieli, 2001; Mrkvicka & Myllymaki, 2023).

As we highlight in the manuscript, the suggested equation is a modification of the Walker-type field significance test and was mainly meant to provide an alternative to FDR control. We thank the reviewer for pointing out that there are specific multiple test corrections designed to account for test dependence and we now refer to these in the revised manuscript. However, we also stress that a review from Cortes et al. (2020) found that the BH procedure outperforms the adaptive procedure by Benjamini & Yekutieli (2001), which is why we focus our analysis on BH.

5. Recommendation to Report Effect Sizes

I recommend adding a recommendation that geoscientists present effect sizes (e.g., physical anomaly magnitude or standardized effect size) alongside p-value maps. Showing effect sizes would mitigate unnecessary debates over whether a statistically significant grid point carries practical or scientific relevance.

We thank the reviewer for this suggestion and have added the recommendation to present effect sizes in Sect. 6.

Minor / Specific Comments

Line 48:

Quote: "In some applications, the resulting FDR threshold may even be higher than the uncorrected threshold."

Comment: As noted in Major Comment 2, this statement is misleading. The threshold is only higher because different nominal significance levels were chosen ($\alpha_{\text{FDR}} = 0.1$

vs. $\alpha_{\text{pointwise}} = 0.05$). Please clarify this sentence to reflect that this is a consequence of parameter choice rather than an intrinsic property of FDR.

We agree, and specified the context.

Lines 84–86:

Quote: "In practice, selecting α_{FDR} approximately twice as large as the target field significance level can lead to corrected thresholds that exceed the nominal local significance level. This may produce the counterintuitive outcome that more grid points appear statistically significant following FDR correction than when using uncorrected local tests."

Comment: Please clarify here as well that this counterintuitive outcome arises strictly from applying different nominal alpha values for local and FDR tests, not from FDR itself.

We agree, and specified the context.

Lines 128–130:

Quote: "The number of stippled gridpoints is even larger than the uncorrected significance testing case (cf. Fig. 1d to Fig. A1)."

Comment: Again, please remind the reader that Fig. 1d uses $\alpha = 0.1$ while Fig. A1 uses $\alpha = 0.05$.

We agree, and specified the context. Furthermore, $\alpha_{\text{FDR}} = 0.1$ and $\alpha_{\text{local}} = 0.05$ are explicitly stated when discussing Fig. 1 in Sect. 3.2.

Lines 156–157:

Quote: "Yet, this behavior only applies to domains dominated by points with relatively large amplitudes, i.e., the FDR correction reduces the significance threshold p^* as expected for domains containing only few grid points with strong anomalies."

Comment: It is not the "amplitude" per se that dictates this behavior, but rather the relative proportion of small p-values (i.e., the overall distribution of test statistics and N). I suggest rephrasing this to focus on p-value proportions rather than physical signal amplitudes.

We thank the reviewer for highlighting this detail and have adjusted the sentence accordingly.

Lines 191–198:

Quote: "For example, one could reduce an initial $\alpha_{in} = 0.1$ to $\alpha_{eff} = \alpha_{in} * \exp(-N_{sig}/N)$..."

Comment: Adjusting nominal alpha post-hoc based on data-derived properties (α_{eff}) is statistically problematic. The significance level alpha must be set a priori; modifying it after observing the fraction of significant tests invalidates standard probabilistic interpretations of the error rate.

We had already acknowledged this drawback in the text (lines 197-198 in the initial version of manuscript): 'We note that the function chosen here to adjust alpha is just an example and may reduce the statistical interpretability of the FDR correction method.'

Nevertheless, to avoid confusion and improve clarity, we have removed the suggestion of a specific α_{eff} , and refer instead to adaptive FDR correction methods in the literature as possible refinements.

Line 218:

Quote: "When a large fraction of grid points exhibit strong anomalies, the resulting FDR threshold can become relatively permissive..."

Comment: Please see Major Comment 1. This permissiveness is a feature of FDR control when signal density is high, rather than a flaw.

We agree that this behaviour is a feature of FDR control, and meant to highlight two points. First, that this property of FDR is seldom acknowledged in the geosciences literature, and on the contrary many studies implicitly or explicitly interpret FDR as automatically leading to stricter thresholds. Second, if users of the FDR correction are not aware of this property, it can lead to potentially overconfident conclusions arising from a gridpoint-wise interpretation of FDR-adjusted tests. We adjusted the sentence to clarify this, in line with the reviewer's suggestion.

References

Cortés, J., Mahecha, M., Reichstein, M., & Brenning, A.: Accounting for multiple testing in the analysis of spatio-temporal environmental data, *Environmental and Ecological Statistics*, 27(2), 293-318, 2020.

Livezey, R. E. and Chen, W. Y.: Statistical Field Significance and its Determination by Monte Carlo Techniques, *Monthly Weather Review*, 111, 46–59, [https://doi.org/10.1175/1520-0493\(1983\)111<0046:SFSaid>2.0.CO;2](https://doi.org/10.1175/1520-0493(1983)111<0046:SFSaid>2.0.CO;2), publisher: American Meteorological Society Section: *Monthly Weather Review*, 1983.

Walker, S. G. T.: Correlation in Seasonal Variations of Weather, III: On the Criterion for the Reality of Relationships Or Periodicities, Meteorological Office, google-Books-ID: cuGtuAAACAAJ, 1914.

Wilks, D. S.: “The Stippling Shows Statistically Significant Grid Points”: How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It, B. Am. Meteorol. Soc., 97, 2263 – 2273, <https://doi.org/10.1175/BAMS-D-15-00267.1>, 2016.